Trialstreamer: Mapping and Browsing Medical Evidence in Real-Time

Benjamin E. Nye
Northeastern University
nye.b@husky.neu.edu

Ani Nenkova
UPenn
nenkova@seas.upenn.edu

Iain J. Marshall
King's College London
iain.marshall@kcl.ac.uk

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Abstract

We introduce *Trialstreamer*, a living database of clinical trial reports. Here we mainly describe the *evidence extraction* component; this extracts from biomedical abstracts key pieces of information that clinicians need when appraising the literature, and also the relations between these. Specifically, the system extracts descriptions of trial participants, the treatments compared in each arm (the *interventions*), and which outcomes were measured. The system then attempts to infer which interventions were reported to work best by determining their relationship with identified trial outcome measures. In addition to summarizing individual trials, these extracted data elements allow automatic synthesis of results across many trials on the same topic. We apply the system at scale to all reports of randomized controlled trials indexed in MEDLINE, powering the automatic generation of *evidence maps*, which provide a global view of the efficacy of different interventions combining data from all relevant clinical trials on a topic. We make all code and models freely available[1] alongside a demonstration of the web interface.[2]

1 Introduction and Motivation

The highest-quality evidence to inform healthcare practice comes from randomized controlled trials (RCTs). The results of the vast majority of these trials are communicated in the form of unstructured text in journal articles. Such results accumulate quickly, with over 100 articles describing RCTs published daily, on average. It is difficult for healthcare providers and patients to make sense of and keep up with this torrent of unstructured literature.

Consider a patient who has been newly diagnosed with diabetes. She would like to consult (in collaboration with her healthcare provider) the available evidence regarding her treatment options. But she may not even be aware of what her treatment options are. Further, she may only care about particular outcomes (for instance, managing her blood pressure). Currently, it is not straightforward to retrieve and browse the evidence pertaining to a given condition, and in particular to ascertain which treatments are best supported for a specific outcome of interest.

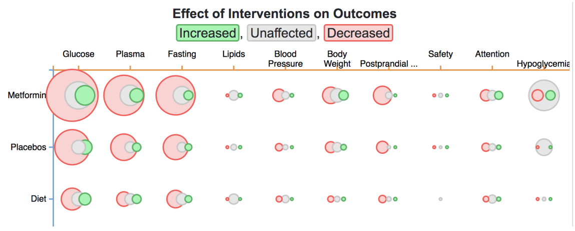


Figure 1: A portion of an example evidence mapping Interventions and their inferred efficacy for Outcomes, given the condition (or Population) of *Type II Diabetes*. These maps are generated automatically using the NLP system we describe in this work.

Trialstreamer is a first attempt to solve this problem, making evidence more browseable via NLP technologies. Figure 1 shows one of the key features of the system: an automatically generated *evidence map* that displays treatments (vertical axis) and outcomes (horizontal) identified for a condition specified by the user (here, migraines). We elaborate on this particular example to illustrate the use of the system in Section 3.

Trialstreamer aims to facilitate efficient evidence mapping with a user friendly method of presenting a search across a broad field (here, being a clinical condition) (Miake-Lye et al., 2016). We use NLP technologies to provide browseable, interactive overviews of large volumes of literature, on-demand. These may then inform subsequent, formal syntheses, or they may simply guide ex-

[1] https://github.com/bepnye/evidence_extraction/
[2] http://bit.ly/trialstreamer

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 63–69
July 5 - July 10, 2020.

ploration of the primary literature. In this work we describe an open-source prototype that enables evidence mapping, using NLP to generate interactive overviews and visualizations of all RCT reports indexed by MEDLINE (and accessible via PubMed).

When mapping the evidence one is generally interested in the following basic questions:

- What interventions and outcomes have been studied for a given condition (population)?
- How much evidence exists, both in terms of the number of trials and the number of participants within these?
- Does the evidence seem to support use of a particular intervention for a given condition?

In the remainder of this paper we describe a prototype system that facilitates interactive exploration and mapping of the evidence base, with an emphasis on answering the above questions. The Trialstreamer mapping interface allows structured search over study populations, interventions/comparators, and outcomes — collectively referred to as PICO elements (Huang et al., 2006). It then displays key clinical attributes automatically extracted from the set of retrieved trials. This is made possible via NLP modules trained on recently released corpora (Nye et al., 2018; Lehman et al., 2019), described below.

2 System Overview

The evidence extraction pipeline is composed of four primary phases. First, text snippets that convey information about the trial's treatments (or *interventions*), outcome measures, and results are extracted from abstracts. Relations between these snippets are then inferred to identify which treatments were compared against each other, and which outcomes were measured for these comparisons. The extracted relations and evidence statements are then used to infer an overall conclusion about the comparative efficacy of the trial's interventions. Finally, the clinical concepts expressed in the extracted spans are normalized to a structured vocabulary in order to ground them in an existing knowledge base and allow for aggregations across trials.

A typical RCT report would pertain to a single clinical condition (the *population*), but might report multiple numerical results, each concerning a particular intervention, comparator, and outcome measure (which we describe as an ICO triplet).

Because the end-to-end task combines NLP subtasks that are supported by different datasets, we collected new development and test sets — 160 abstracts in all, exhaustively annotated — in order to evaluate the overall performance of our system. Two medical doctors[3] annotated these documents with the all of the expressed entities, their mentions in the text, the relations between them, the conclusions reported for each ICO triplet and the sentence that contains the supporting evidence for this (Lehman et al., 2019).

We were unable to obtain normalized concept labels for the ICO triplets due to the excessive difficulty of the task for the annotators.

Modeling decisions were informed by the 60 document development set, and we present evaluations of the first four information extraction modules with regard to the 100 documents in the unseen test set.

2.1 Preprocessing

Enabling search over RCT reports requires first compiling and indexing all such studies. This is, perhaps surprisingly, non-trivial. One may rely on "Publication Type" (PT) tags that codify study designs of articles, but these are manually applied by staff at the National Library of Medicine. Consequently, there is a lag between when a new study is published and when a PT tag is applied. Relying on these tags may thus hinder access to the most up-to-date evidence available. Therefore, we instead use an automated tagging system that uses machine learning to classify articles as RCT reports (or not). This model has been validated extensively in prior work (Marshall et al., 2018), and we do not describe it further here.

Next, we replace all abbreviations with their long forms using the Ab3P algorithm (Sohn et al., 2008). Using long forms has the complementary advantages of improving PICO labeling accuracy while also reducing the amount of context needed for prediction by downstream model components.

2.2 Study Descriptor Recognition

PICO Elements

In order to identify the spans of text corresponding to the PICO elements of the trial, we use the *EBM-NLP* corpus (Nye et al., 2018). This is a dataset

[3] Hired via Upwork (http://www.upwork.com).

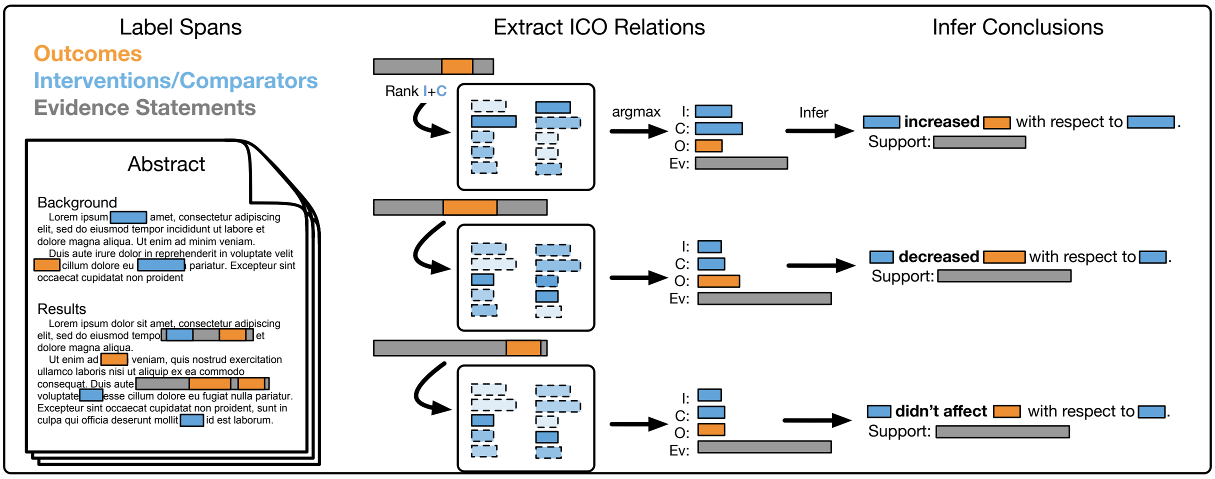

Label Spans
Outcomes
Interventions/Comparators
Evidence Statements
Abstract
Background
Lorem ipsum amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam.
Duis aute irure dolor in reprehenderit in voluptate velit cillum dolore eu pariatur. Excepteur sint occaecat cupidatat non proident
Results
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempo et dolore magna aliqua.
Ut enim ad veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute voluptate esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit id est laborum.
Extract ICO Relations
Rank I+C
argmax
I:
C:
O:
Ev:
Infer
Infer Conclusions
increased with respect to .
Support:
decreased with respect to .
Support:
didn't affect with respect to .
Support:

Figure 2: Overview of the evidence extraction pipeline, applied to all RCT article abstracts automatically identified. Text spans are first extracted from these abstracts, then assembled into relations that reflect the structure of the trials, and finally used to infer the effect interventions were reported to have on measured outcomes, as compared to the control treatment.

	F1	Precision	Recall
Tokens	0.63	0.56	0.72
Clinical Entities	0.67	0.55	0.87

Table 1: Macro-averaged scores for ICO span prediction at both the token and clinical entity level.

	F1	Precision	Recall
Evidence	0.69	0.53	0.97

Table 2: Performance for identifying evidence-bearing sentences.

comprising ∼5,000 abstracts of RCT reports that have been annotated to demarcate textual spans that describe the respective PICO elements. In addition to these spans, it contains more granular annotations on information within spans (e.g., specific Population attributes like age and sex).

We follow our prior work (Nye et al., 2018) in training a BiLSTM-CRF model that learns to jointly predict each PICO element using EBM-NLP. Recent work has shown the efficacy of BERT (Devlin et al., 2018) representations in this space, e.g., Beltagy *et al.* achieved state-of-the-art performance on EBM-NLP using this approach (2019). Therefore, for all text encoding we use BioBERT (Lee et al., 2019), which was pretrained on PubMed documents.[4]

Results for Interventions/Comparators and Outcomes on our test set are reported in Table 1. Since these spans will serve as inputs to downstream models in the pipeline, high recall at the expense of precision is preferable; we will allow subsequent classifiers to discard spurious spans. We achieve 0.87 recall at the clinical concept level.

Evidence Statements

In addition to PICO elements, we extract all sentences in the abstract that are predicted to contain evidence concerning the relative efficacy of an Intervention. Our training data for this model is sourced from the *Evidence-Inference* corpus (Lehman et al., 2019), which comprises ∼10,000 annotated 'prompts' across ∼2,400 unique full-text articles. Each prompt specifies an Intervention, a Comparator, and an Outcome. Doctors have annotated the prompts for each article, supplying an extracted snippet that presents the conclusion for these ICO elements, as well as an inference concerning whether the Outcome increased, decreased, or remained the same in the intervention group (as compared to the comparator group).

We frame evidence identification as a sentence classication task, and train a linear classification layer on top of BioBERT outputs. Our positive training examples are the sentences containing evidence snippets in Evidence-Inference, and we draw an equal number of length-matched negatives randomly from the rest of the document. As shown in Table 2, we achieve extremely high recall on the test set, but only middling precision. On inspection, many of these false positives are sentences from the conclusion that provide a high-

[4] For PICO tagging on EBM-NLP we found that BioBERT performed comparably to SciBERT (Beltagy et al., 2019).

level summary of the evidence, but aren't the best evidence statement — as provided by the annotator — for any given ICO prompt.

2.3 Relation Extraction

To transform the extracted spans into a semantic representation of the trial that can be used to construct an evidence map, we must identify all instances of an outcome being reported, and infer which two treatments were being directly compared as the intervention and comparator with respect to said outcome. Finally, given each assembled ICO prompt, we can then predict the trial's findings regarding whether the outcome increased, decreased, or was not statistically different under the intervention versus the comparator. In effect, we are aiming to jointly extract ICO prompts *and* infer the directionality of the results reported concerning these, whereas prior work (Nye et al., 2018; Lehman et al., 2019) has considered these problems only in isolation.

Our strategy for assembling ICO prompts is informed by the style in which results are commonly described in abstracts.When results are described in an article the outcome is typically referenced explicitly, while the intervention and especially the comparator are often referenced either indirectly ("Mean headache duration was similar between groups"), or not at all ("No significant difference was observed for recovery time"). In the fully annotated dev set collected for this work, 87% of outcomes were described explicitly in an evidence span, while only 28% of treatments were explicit.

Motivated by this observation, we use the (explicit) outcomes extracted from an evidence snippet as a starting point; for each of these outcomes, the associated intervention and comparator are then inferred. This has the significant advantage of explicitly linking each outcome to the evidence that will be used to infer the directionality of the reported finding. This also provides the end-user with an interpretable rationale for the inference concerning treatment efficacy.

To link candidate extracted treatments to specific outcome mentions, we train a model that takes in a candidate treatment, an evidence statement containing the outcome, and the surrounding context from the document, and predicts if the treatment is the participating intervention, the participating comparator, or if it is not involved in this particular evaluation. We use the evidence-inference corpus to provide training examples for the first two classes, and manually generate negative samples for the final class. The negatives are constructed to mimic common errors that the treatment extraction module made on the dev set, including: mislabeling an outcome as a treatment; extracting compound phrases containing multiple individual treatments; and, finally, extracting spurious spans that don't represent a study descriptor.

	F1	Precision	Recall
No Difference	0.91	0.94	0.89
Increased	0.73	0.69	0.77
Decreased	0.76	0.75	0.78

Table 3: Per-class prediction scores for each outcome in the test set.

The model is a linear classifier on top of BioBERT. Inputs are constructed as: [CLS] TREATMENT [SEP] EVIDENCE. CONTEXT. [SEP]. We experimented with different slices of the document as the context, and achieved the highest dev performance using the first four sentences of the article. The class probabilities from this model are used to rank the possible interventions and comparators for each outcome, and when sufficiently probable candidates are identified we generate a complete ICO prompt.

After assembling all ICO prompts in a document, we feed them to a final classifier to predict the directionality of findings for each outcome, with respect to the given intervention and comparator. This model is trained over the evidence-inference corpus using the provided I, C, and O spans coupled with the sentences that contain the corresponding evidence statement. Empirically, we found that signal for the classifier is dominated by the outcome text and evidence span, with almost no contribution from the intervention and comparator. This is unsurprising given the regularity of the language used to describe conclusions. The reported directionality of the result is almost exclusively framed with respect to the intervention, and only 4.0% of all outcomes ever have different results for another I+C linking within the same document. The best performing model input was simply [CLS] OUTCOME [SEP] EVIDENCE [SEP], and the results on the test set are reported in Table 3.

Strict	Precision	Recall
Extracted spans	0.26	0.24
Expert spans	0.23	0.26
Relaxed	Precision	Recall
Extracted spans	0.32	0.34
Expert spans	0.31	0.34

Table 4: Performance for predicting an article's exact MeSH terms using the rule-base system, run on both the automatically extracted spans and the expert-provided test spans.

False Neg		False Pos	
Name / Count		**Name / Count**	
Humans	185	Patients	115
Middle Aged	93	Aging	42
Adult	84	Therapeutics	42
Aged	62	Weights/Measures	33
Double-Blind Method	50	Placebos	33
Treatment Outcome	42	Time	21
Adolescent	39	Serum	17
Prospective Studies	27	Safety	17
Time Factors	20	Pain	16
Child, Preschool	20	Women	14

Table 5: Ten most common over- and under-predicted MeSH terms for the test set of 191 articles.

2.4 Normalizing PICO Terms

In order to standardize the language used to categorize the articles with respect to their PICO elements, we turn to the structured vocabulary provided by the National Libaray of Medicine (NLM) in the form of Medical Subject Heading (MeSH) terms. This resource codifies a comprehensive set of medical concepts into an ontology that includes their descriptions, properties, and the structured relationships between them. Each article in the MEDLINE database maintained by the NLM is annotated with the relevant MeSH terms by expert library scientists (subject to the same lag that necessitates an RCT classifier instead of relying on annotated Publication Types).

To induce relevant MeSH terms for an extracted text span, we reproduced the method described in the Metamap Lite paper (Demner-Fushman et al., 2017) to extract MeSH terms describing the PICO elements. In short, we generated a large dictionary of synonyms for medical terms algorithmically using data from the UMLS Metathesaurus, with synonyms being matched to unique identifiers pertaining to concepts in the MeSH vocabulary. We used this dictionary to map matching strings in our extracted PICO text to MeSH terms, yielding a set of normalized concepts describing each of the population, intervention, and outcome spans in the documents.

To evaluate the accuracy of this approach, we compare the differences between the MeSH terms produced by our system against those provided by the NLM for the 191 articles that comprise the test set for EBM-NLP.

The test articles are provided with an average of 14.8 MeSH terms per article, while our system induces 14.0 terms on average. The strictest evaluation for this module is to require exact matches between the predicted MeSH terms and the official MEDLINE terms – a daunting task given the $30,000$ possible labels we have to chose from. However, because the concepts in the ontology exist in varying levels of specificity (for example *Migraine with Aura* is a subset of *Migraine Disorders*), it is often the case that the predicted MeSH term is sufficiently close to the provided MeSH term for practical purposes, but differs in the level of specificity.

To better characterize the performance of our approach, we therefore also consider relaxing the equivalence criteria to include matching immediate parents or children in the MeSH hierarchy. This modification results in a 42% relative increase in recall and a 23% increase in precision, as shown in Table 4.

We observe that while the absolute accuracy is not high, this technique generally captures the key terms for the PICO elements. The most common mistakes, shown in Table 5, mostly involve missing age or publication type terms, and systematic differences between the general MeSH terms commonly applied to articles (for example, we might apply *Patients* rather than *Humans*).

A more sophisticated aligment between the way MeSH terms are applied by experts and the terms produced by our system has the potential to improve the overall effectiveness of the tool; we intend to pursue this in future work.

3 Illustrative Example

To illustrate the envisioned use of our automatic mapping system, we return to the example we began with at the outset of this paper: seeking evidence concerning treatment of Type II Diabetes. To begin, the user specifies a condition (Population) of interest. We rely on Medical Subject Headings (MeSH) terms,[5] which as dis-

[5] https://www.ncbi.nlm.nih.gov/mesh

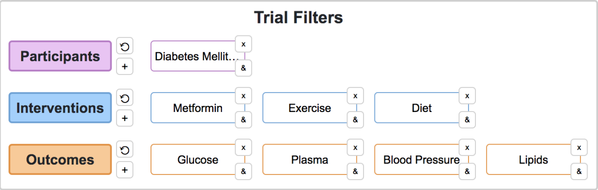
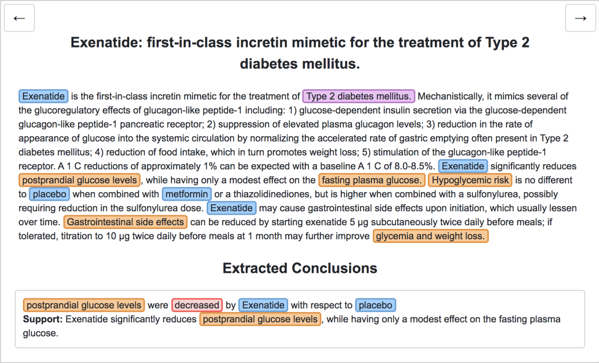


Figure 3: View of a collected set of concepts used to specify trials of interest. The search interface allows concepts to be combined using and/or operators.



Exenatide: first-in-class incretin mimetic for the treatment of Type 2 diabetes mellitus.

Exenatide is the first-in-class incretin mimetic for the treatment of Type 2 diabetes mellitus. Mechanistically, it mimics several of the glucoregulatory effects of glucagon-like peptide-1 including: 1) glucose-dependent insulin secretion via the glucose-dependent glucagon-like peptide-1 pancreatic receptor; 2) suppression of elevated plasma glucagon levels; 3) reduction in the rate of appearance of glucose into the systemic circulation by normalizing the accelerated rate of gastric emptying often present in Type 2 diabetes mellitus; 4) reduction of food intake, which in turn promotes weight loss; 5) stimulation of the glucagon-like peptide-1 receptor. A 1 C reductions of approximately 1% can be expected with a baseline A 1 C of 8.0-8.5%. Exenatide significantly reduces postprandial glucose levels, while having only a modest effect on the fasting plasma glucose. Hypoglycemic risk is no different to placebo when combined with metformin or a thiazolidinediones, but is higher when combined with a sulfonylurea, possibly requiring reduction in the sulfonylurea dose. Exenatide may cause gastrointestinal side effects upon initiation, which usually lessen over time. Gastrointestinal side effects can be reduced by starting exenatide 5 μg subcutaneously twice daily before meals; if tolerated, titration to 10 μg twice daily before meals at 1 month may further improve glycemia and weight loss.

Extracted Conclusions

postprandial glucose levels were decreased by Exenatide with respect to placebo
Support: Exenatide significantly reduces postprandial glucose levels, while having only a modest effect on the fasting plasma glucose.

Figure 4: Detailed view of selected abstracts that contribute to the evidence map. These are automatically annotated with all extracted information.

cussed above is a structured vocabularly maintained by the NLM. We allow users to enter a search string and provide auto-complete options from the MeSH vocabulary. Users can additionally provide interventions or outcomes of interest to further narrow the search. We show an example of a constructed set of filters in Figure 3.

Once a set of search terms is specified, relevant RCTs are retrieved from the comprehensive and up-to-date database.[6] The interface then displays counts of unique interventions and outcomes covered by the retrieved trials. Each bar in these plots can be clicked to explicitly include that concept in the search terms, allowing for a data-driven approach to building up the search parameters via iterative refinement.

At this point, the evidence map shown in Figure 1 is also displayed, providing a summary of the evidence available for the effectiveness of the selected interventions with respect to their co-occurring outcomes. The user can mouse-over plot elements to view tooltips that include snippets of contributing evidence from the underlying abstracts, or click through to browse these texts annotated with all of the extracted information, as shown in Figure 4.

4 User Study

To evaluate the system's utility for a real-world task, we provided the tool to a team of researchers at Cures Within Reach for Cancer (CWR4C).[7] Domain experts reviewed the extracted ICO conclusions and automatically generated plots for a randomly selected subset of documents pertaining to cancer trials, a domain that is particularly challenging given the prevalence of complex compound interventions that often share individual components between trial arms.

The reviewers were asked to evaluate the types of mistakes made by the system as well as the overall precision and recall of the extracted conclusions for each document. Across 21 documents average precision was 54% and average recall was 75%, and the team expressed excitement about the efficacy of the system for their purposes. CWR4C has continued to work with this tool as a source of information about cancer-related clinical trials.

5 Conclusions

We have presented the evidence extraction component of Trialstreamer, an open-source prototype that performs end-to-end identification of published RCT reports, extracts key elements from the texts (intervention and outcomes descriptions), and performs relation extraction between these, i.e., attempts to determine which intervention was reported to work for which outcomes.

We use this pipeline to provide fast, on-demand overviews of all published evidence pertaining to a condition of interest. Moving forward, we hope to refine the linking of extracted snippets to structured vocabularies to run a more comprehensive user-study to evaluate the use of the system in practice by different types of users. We also hope to develop a joint extraction and inference model, rather than relying on the current pipelined approach.

Acknowledgements

This work was funded in part by the National Institutes of Health (NIH) under the National Library of Medicine (NLM) grant 2R01LM012086, and by the National Science Foundation (NSF) CAREER award 1750978.

[6] We update this database nightly by scanning MEDLINE for new RCT reports using our RCT classifier (Marshall et al., 2018).

[7] https://www.cwr4c.org/

References

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*.

Dina Demner-Fushman, Willie J Rogers, and Alan R Aronson. 2017. Metamap lite: an evaluation of a new java implementation of metamap. *Journal of the American Medical Informatics Association*, 24(4):841–844.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. 2006. Evaluation of pico as a knowledge representation for clinical questions. In *AMIA annual symposium proceedings*, volume 2006, page 359. American Medical Informatics Association.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring Which Medical Treatments Work from Reports of Clinical Trials. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 3705–3717.

Iain J Marshall, Anna Noel-Storr, Joël Kuiper, James Thomas, and Byron C Wallace. 2018. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. *Research synthesis methods*, 9(4):602–614.

Isomi M Miake-Lye, Susanne Hempel, Roberta Shanman, and Paul G Shekelle. 2016. What is an evidence map? a systematic review of published evidence maps and their definitions, methods, and products. *Syst. Rev.*, 5:28.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access.

Sunghwan Sohn, Donald C Comeau, Won Kim, and W John Wilbur. 2008. Abbreviation definition identification based on automatic precision estimates. *BMC bioinformatics*, 9(1):402.